AD612616

QUALITY CONTROL AND RELIABILITY

TECHNICAL REPORT TR 5

STATISTICAL THEORY AND METHODS FOR VALIDATING RESULTS OF SAMPLING INSPECTION BY ATTRIBUTES

COPY 2 OF 3 42p
HARD COPY $. 2.00
MICROFICHE $. 0.50

DDC
RECEIVED
MAR 29 1965
TISIA E

16 APRIL 1962
OFFICE OF THE ASSISTANT SECRETARY OF DEFENSE
(INSTALLATIONS AND LOGISTICS)
WASHINGTON 25, D. C.

ARCHIVE COPY

Best
Available
Copy

OFFICE OF THE ASSISTANT SECRETARY OF DEFENSE
WASHINGTON 25, D.C.

ISTALLATIONS AND LOGISTICS

16 April 1962

Statistical Theory and Methods for Validating Results of Sampling Inspection by Attributes TR-5

Quality Control & Reliability

This technical report was prepared on behalf of the Office of the Assistant Secretary of Defense (Installations and Logistics) by the Chemical Corps, Department of the Army. It was written by Mr. Henry Ellner of the Chemical Corps Materiel Command, Edgewood, Maryland.

This publication provides the statistical theory and techniques underlying the procedures and tables furnished in DoD Handbook H109, "Statistical Procedures for Determining Validity of Suppliers' Attributes Inspection," dated 6 May 1960.

TABLE OF CONTENTS

APPENDIX

STATISTICAL THEORY AND METHODS FOR VALIDATING RESULTS OF SAMPLING INSPECTION BY ATTRIBUTES

Henry Ellner
Chemical Corps Materiel Command, Department of the Army

Simple but robust statistical methods are described and developed for use in validating suppliers' inspection records of attribute sampling data. The methods are essentially two-sample significance tests for homogeneity of discrete variates treated as continuous, and the combination of their probabilities to test the hypothesis of over-all agreement of paired inspection results. The statistical theory and techniques presented in this paper form the basis for DoD Handbook H109, "Statistical Procedures for Determining Validity of Suppliers' Attributes Inspection." The procedures of the handbook constitute a system of product verification inspection wherein the consumer's sampling results establish the reliability of the supplier's acceptance sampling records and provide independent estimates of the quality of product submitted for acceptance. The characteristics and application of alternate tests are discussed. Utilization of the power function of the significance tests affords substantial reduction in the amount of product verification inspection.

1. INTRODUCTION. As an extension of the principle long recognized by industry [1] that amount of inspection is a function of control of product quality, the Department of Defense established a policy in April 1954 that optimum use be made of inspection data obtained by suppliers in determining acceptability of supplies. This broad policy was implemented by prescribing uniform procedures in military and

federal series of specifications requiring the supplier to perform examinations and tests itemized in the quality assurance provisions and to maintain records of his inspection results. Verification of the supplier's compliance with technical requirements of the contract was made the responsibility of the Government representative.

When a system of sampling inspection, like MIL-STD 105 (Sampling Procedures and Tables for Inspection by Attributes), is required of the supplier, there is an incentive for him to upgrade and maintain an acceptable quality level for the product submitted for consumer's acceptance. The consumer's product inspection then can be adjusted to an amount necessary to verify the sampling results recorded by the supplier.

To assist the Government representative in establishing the validity of the supplier's inspection records, the Office of the Assistant Secretary of Defense (Supply and Logistics) has published Quality Control and Reliability Handbook (Interim) H109 [2] , which provides procedures for validating the results of sampling inspection by attributes as recorded by a supplier. The underlying mathematical and statistical principles of these procedures are included in this paper. The basic concepts were derived from scattered literature sources and adapted to meet the exigencies of the field inspector. For this purpose, simple approximate tests for variates from discrete distributions were investigated and developed to provide a systematic approach for accomplishing product verification inspection.

2. PRODUCT VERIFICATION INSPECTION. To make the discussion more concrete, reference will be made to MIL-STD 105 [3] and the terms

defined in that document will be used in what follows. In inspection by attributes, the unit of product is classified simply as defective or effective (nondefective) with respect to a given requirement or a set of requirements. The requirement may be an individual checkpoint and the set may be a group of characteristics of equal importance listed under a single acceptable quality level (AQL) in the specification. We shall assume that even when a measurement along a continuous numerical scale is possible, such measurement will be classified as conforming or nonconforming with the tolerance limits prescribed.

Let us now suppose that a supplier has drawn a single sample from an inspection lot in accordance with MIL-STD 105, and has noted the number of conforming ana nonconforming items in the sample. The consumer has proceeded likewise by selecting an equal or smaller sample from the same lot (the size of the sample will be adjusted in subsequent trials). We shall assume that the lot size is large relative to the total sample size (say, at least 8:1) so that the respective samples can be considered as independently drawn from a binomial population. When the lot size i relatively small, the condition is imposed that the samples be drawn without replacement. The results of the two inspections are denoted symbolically in a 2 X 2 table as below:

TABLE 1

Notation for Two-Sample Test for Homogeneity

	Defective	Effective	Sample Size
Supplier's Sample	d_s	$n_s - d_s$	n_s
Consumer's Sample	d_c	$n_c - d_c$	n_c
Total	d_t	$n_t - d_t$	n_t

The sample sizes of the supplier and the consumer are represented by n_s and n_c, and their total by n_t. The number of defectives recorded by the supplier and consumer are symbolized by d_s and d_c, and their sum by d_t. Product verification inspection is accomplished by comparing the proportion defective in the supplier's sample with the proportion defective in the consumer's sample. The comparison is considered a test of homogeneity of the two samples since the concern is whether the fraction defectives observed would be such as would only occur by chance selection of the sample units, inspection being uniformly performed. The problem is to set up criteria so that discrepancies arising by chance alone are differentiated from those generated by disparities in the inspection practice. Statistically this can be accomplished by significance tests for homogeneity of the two sample results.

3. SIGNIFICANCE TESTS FOR HOMOGENEITY. A common test of significance for dichotomized data is the chi-square test [4] and equivalent alternates. When the expected number of defectives is small, say less than five, Fisher's exact test ([5] , Section 21.02) is generally advised. For routine testing, these techniques all involve extensive computation, and consequently are not suitable for verification purposes. Short cut procedures [6, 7, 8, 9] devised to meet this problem, including nomograms and extensive tabulations of Fisher's exact test, are likewise wanting in that multiple entries are necessary or that tables required are too lengthy and numerous.

A test for homogeneity, applicable when the proportion of defectives d_t/n_t is small, say 0.20 or less, is one which compares samples from populations approximated by the Poisson type of distribution. Przyborowski & Wilenski [10] considered two observations (in our notation: d_c and d_s) originating from two Poisson-distributed populations with unknown means, and for the symmetrical case $n_c = n_s$ they proposed an "exact" test for the equality of these means. Barnard [11] extended their method to the case $n_c \neq n_s$ reducing the procedure to a simple test for the variance - ratio F. Bross and Kasten [12] derived an equivalent technique for the case $n_c \neq n_s$ and published charts for avoiding or reducing computations. Cox [13] proposed a variance-ratio test, treating variates as continuous, for the equivalence of two Poisson processes. David and Johnson [14] and Lancaster [15] suggested a probability integral transformation when the variable is discontinuous, which was further amplified by Lancaster [16] who proposed the use of the mid or median probability as a test function for discrete distributions. The apparently different tests for Poisson variates can be shown to be essentially equivalent when the number of events observed are not too small. This has been noted by Barton [17] and is further discussed in the development which follows. It will be shown that Cox's method has certain properties which make it preferable for use in product verification inspection. For accurate results in very small samples, this method is to be preferred to approximate chi-square methods. For larger samples, when appropriate tables of critical values are not available, the approximate chi-square methods will be found to be surprisingly good. Probabilities are obtained which correspond closely to those given by the respective test functions suggested by Cox and Lancaster.

4. "EXACT" TESTS OF SAMPLES FROM TWO POISSON SERIES. Before the features of the test functions of Cox and Lancaster can be discussed, it will be necessary to derive the "exact" test for comparing two Poisson-distributed observations. Suppose d_s and d_c of Table 1 approximately follow independent Poisson distributions so that:

$$(1)\quad P(d_s,d_c \,|\, p'_s,p'_c) = P(d_s)\ .\ P(d_c) = \frac{e^{-p'_s n_s}(p'_s n_s)^{d_s}}{d_s!} \cdot \frac{e^{-p'_c n_c}(p'_c n_c)^{d_c}}{d_c!}\ ,$$

where:

p'_s = the expected proportion defective in the supplier's sample n_s ,

p'_c = the expected proportion defective in the consumer's sample n_c.

Under the null hypothesis $p'_s = p'_c = p'_o$ so that Equation (1) reduces to:

$$(2)\quad P(d_s,d_c \,|\, p'_o) = \frac{e^{-p'_o(n_s + n_c)}(p'_o)^{d_t}\, n_s^{\ d_s}\, n_c^{\ d_c}}{d_s!\ d_c!}\ ,$$

which can be rewritten as:

$$(3)\quad P(d_s,d_c \,|\, p'_o) = P(d_c \,|\, d_t)\ P(d_t \,|\, p'_o)$$

$$= \frac{d_t!\quad n_s^{\ d_s}\qquad n_c^{\ d_c}}{d_s!\ d_c!\ (n_s + n_c)^{d_s}\ (n_s + n_c)^{d_c}} \cdot \frac{e^{-p'_o(n_s + n_c)}(p'_o n_s + p'_o n_c)^{d_t}}{d_t!}\ .$$

But we need the probability of getting some pair of results having the same total $d_s + d_c = d_t$; and so the relative probability, on the null hypothesis, of getting the pair (d_s,d_c) out of all results with the same total d_t is:

$$(4)\quad P(d_c \,|\, d_t) = \frac{P(d_c \,|\, d_t)\ P(d_t \,|\, p'_o)}{P(d_t \,|\, p'_o)}$$

$$= \frac{d_t!}{d_s!\,d_c!}\ \left(\frac{n_c}{n_s + n_c}\right)^{d_c}\left(\frac{n_s}{n_s + n_c}\right)^{d_s}\ .$$

If we let $r = \frac{n_s}{n_c}$ then:

$$(5) \quad P(d_c \mid d_t) = \frac{d_t!}{d_s! d_c!} \left(\frac{1}{1 + r}\right)^{d_c} \left(\frac{r}{1 + r}\right)^{d_s} \quad .$$

We note that conditionally on d_t, d_c is binomially distributed with parameters, $\frac{1}{1 + r}$ and d_t, which can be used as the basis for a significance test. Accordingly:

$$(6) \quad F(y) = \sum_{y = d_c}^{d_t} \binom{d_t}{y} \left(\frac{1}{1 + r}\right)^{v} \left(\frac{r}{1 + r}\right)^{d_t - y} = I_{\frac{1}{1 + r}} (d_c, d_s + 1),$$

where $I_x (p,q)$ is the incomplete β - function representation of a sum of binomial probabilities.

If the only admissible alternative to the null hypothesis $p_s' = p_c' = p_o'$ is $p_c' > p_s'$ then the appropriate critical region, in the Neyman-Pearson sense, for rejection of the null hypothesis is defined by

$d_s \leq k_1 (d_t, \alpha)$ or $d_c \geq k_2 (d_t, \alpha)$,

where α is the risk of the first kind of error and where

$$(7) \quad P \left\{ d_c \geq k_2 (d_t, \alpha) \mid d_t, p_s' = p_c' \right\} \leq \alpha \, .$$

For the "exact" test this may be expressed by:

$$(8) \quad I_{\frac{1}{1 + r}} (d_c, d_s + 1) \leq \alpha \quad .$$

This inequality may be written in terms of the probability distribution function $P_{f_1, f_2} (F)$ of the F distribution with (f_1, f_2) degrees of freedom since:

$$P_{f_1, f_2}(F) = I_x (p,q)$$

where $f_1 = 2q$, $f_2 = 2p$ and $F = \frac{p}{q} \frac{1 - x}{x}$ with the result that

$$(9) \quad P_{2d_s + 2, 2d_c} \left(\frac{r\, d_c}{d_s + 1}\right) \leq \alpha \, .$$

Inequalities (8) and (9) establish a level of significance which does not exceed α . The true level of significance depends upon the unknown p_0' and may in some cases for small (d_s, d_c) be considerably less than α .

5. "APPROXIMATE" TESTS FOR POISSON VARIATES. In inverse binomial sampling, with d fixed, Barnard [18] pointed out that when p' is small and n is large the number of sample items drawn in sequence up to the d^{th} event is distributed approximately as $(2p')^{-1} \chi^2_{2d}$, where χ^2_{2d} denotes a chi-square variate with 2d degrees of freedom and p' represents the true rate. For direct binomial sampling, approximated by Poisson's exponential binomial limit, in which the number of events d occurring in a fixed n is observed, we have

$$(10) \quad P(x \geq d) = \sum_{x = d}^{\infty} \frac{e^{-p'n} (p'n)^x}{x!} = P\left(\frac{1}{2p'} \chi^2_{2d} \leq n\right), \text{ and}$$

$$(11) \quad P(x \geq d + 1) = P\left(\frac{1}{2p'} \chi^2_{2d + 2} \leq n\right).$$

Cox [13] suggested an approximation to $P(x > d)$ in which d is treated as a continuous variate by taking a quantity intermediate between (10) and (11):

$$(12) \quad P(x > d) \simeq P\left(\frac{1}{2p'} \chi^2_{2d + 1} \leq n\right),$$

which implies that probabilities are calculated as if

(13) $2p'n$ is distributed as $\chi^2_{2d + 1}$.

When two populations with proportions defective p_s', p_c' are compared by means of samples n_s, n_c which exhibit d_s, d_c defectives, then, from (12) we compute the ratio:

$$(14) \quad \frac{2p_s' n_s}{2d_s + 1} \div \frac{2p_c' n_c}{2d_c + 1}$$

which is distributed approximately as F with $(2d_s + 1, 2d_c + 1)$ degrees of freedom. Thus, we may test the hypothesis that $p'_s = p'_c = p'_o$ against the alternate hypothesis that $p'_c > p'_s$ by referring

$$(15)\quad F = r\,\frac{d_c + 0.5}{d_s + 0.5}$$

to the F tables with $(2d_s + 1, 2d_c + 1)$ degrees of freedom for the appropriate α percent point.

This may be represented by

$$(16)\quad P_{2d_s + 1, 2d_c + 1}\left(r\,\frac{d_c + 0.5}{d_s + 0.5}\right) \le \alpha \text{ , or}$$

$$(17)\quad I_{\frac{1}{1 + r}}\,(d_c + 0.5,\ d_s + 0.5) \le \alpha \ .$$

It is now clear that the "exact" tests given by (8) and (9) have been modified slightly to yield the approximate tests of (16)and (17). The modification has the effect of making the true level of significance less dependent upon the unknown p'_o and to approximate the nominal value of α when averaged over d_t.

6. ALTERNATIVE "APPROXIMATE" TESTS FOR POISSON VARIATES. The median probability defined by

$$(18)\quad \tfrac{1}{2}\left\{P(d_c \mid d_t) + P(d_c + 1 \mid d_t)\right\}$$

was considered by Lancaster [19] as a test function for paired Poisson variates. In terms of the incomplete β- function the critical region may be expressed by:

$$(19)\quad \tfrac{1}{2}\left\{I_{\frac{1}{1 + r}}\,(d_c, d_s + 1) + I_{\frac{1}{1 + r}}\,(d_c + 1, d_s)\right\} \le \alpha \ .$$

A comparison of the critical regions expressed by (8), (17) and (19) shows that for the rejection rule:

$$d_c \ge k_2\,(d_t,\ \alpha\,),$$

the critical values of the medium probability test are bounded by the corresponding values for the other two tests.

Using the median probability test as a basis for comparison, Lancaster [19] has shown arithmetically that the median probability usually closely corresponds with the probability of the uncorrected chi-square test for sets of two counts, x_1 and x_2. His investigation was limited to the simple form:

$$(20) \quad X^2 = (x_1 - x_2)^2 \,/\, x_1 + x_2 \quad .$$

A similar test may be developed for the paired variates d_s and d_c, of Table 1, which we shall suppose to be Poisson distributed. Accordingly, for a one-sided test we have:

$$(21) \quad X^2 = \frac{\left(d_c - \frac{d_t}{1 + r}\right)^2}{\frac{d_t}{1 + r}} + \frac{\left(\frac{r\, d_t}{1 + r} - d_s\right)^2}{\frac{r\, d_t}{1 + r}} \quad ,$$

where, as before, $r = n_s/n_c$ and $d_t = d_c + d_s$. Equation (21) reduces to:

$$(22) \quad X^2 = (r\, d_c - d_s)^2 \,/ r\, d_t \; , \quad \text{or}$$

$$(23) \quad X = \frac{r\, d_c - d_s}{(r\, d_t)^{\frac{1}{2}}} \quad ,$$

where X is approximately normally distributed. When d_t is not too small and r is not much larger than one, the χ approximation yields probabilities which correspond closely to the median probabilities. As d_t increases r may increase without essentially disturbing the correspondence of the probabilities of the two alternate "approximate" tests.

7. <u>POWER FUNCTION OF TESTS FOR POISSON VARIATES</u>. The Neyman-Pearson theory of tests conside's all tests of the same size and lays down objective standards for selecting the best test. The theory introduces the

term, "power of a test," relative to the alternate hypothesis, to denote the probability of correctly rejecting the null hypothesis when an alternative is true. Of all tests at a given significance level, the most preferred is the one which has the maximum power relative to all the alternate hypotheses considered. The probability of rejecting the null hypothesis H_0, regarded as a function of H', where H' is any of the admissible alternates to H_0, is called the power function of the test. If we commence with the determination of the critical region subject to (7) we can calculate the power function of a given test of significance. Thus, for the "exact" test all points satisfying (8) are entered in (1) and the absolute probabilities are summed. Similarly, for the "approximate" test all points satisfying (17) are entered in (1) for addition of the absolute probabilities. Tables 2 and 3 provide the actual probabilities associated with the respective one-sided tests of the null hypothesis $p'_c = p'_s$ against the alternatives $p'_c = 2\ p'_s$, $p'_c = 3\ p'_s$ and $p'_c = 4.5\ p'_s$ for $r = 1, 2, 3, 5$ and 8, respectively, over a range of nuisance parameters $p'_s n'_s$, which may be encountered in practice. A similar table can be computed for all critical values of the alternate "approximate" test satisfying (19). The size of this test and its power function are somewhat less than those of the "approximate" test (17).

The arrangement of Table 2 and Table 3 clearly reveals that the true significance level is a function of the expected number of defectives in the supplier's sample and the ratio of the supplier's sample size to the size of the consumer's validation sample. For the "exact" test, under the null hypothesis, $p'_c = p'_s$, the size of the test increases on the average by a factor of ten as $p'_s n'_s$ increases from 0.75 to 12.00. In contrast, for the "approximate" test, α increases about 1.5 times over the same range of expected number of defects. Furthermore, the mean level of significance of the entries summed over the five tabular values of r for the "exact" and "approximate" tests are 0.017 and 0.051, respectively. The conclusion is that the "approximate" test more effectively controls the size of the test at the significance level of 0.05 than the "exact" test.

Since we can generally estimate $p_s'n_s$ from the supplier's record of inspection results and the AQL under which he is operating, we can select the power of test by adjusting the sample size ratio r commensurate with the relative fraction defective, p_c'/p_s', which should be detected if it exists. This power can be further augmented by simple pooling of inspection results for a given r until the expected number of defectives for the supplier's samples exceeds the desired values of $p_s'n_s$. A. Birnbaum [20] has considered various methods of comparing two Poisson processes in terms of the ratio of their parameters, and suggests for fixed samples an accumulation of observations until the total number of defectives d_t is sufficient to yield the power of test desired.

The values of $p_s'n_s$ and r in Table 3 were selected so that corresponding power function curves could be obtained for a set of alternate hypotheses. Thus, the probabilities of rejection associated with $p_c'/p_s' = 1, 2, 3$ and 4.5, respectively, and subject to the parameters $p_s'n_s = 1.50$ and $r = 1$, approximate the rejection rates tabulated for the following row and columnar headings of Table 3:

$p_s'n_s = 2.25$ and $r = 2$; $p_s'n_s = 3.00$ and $r = 3$;

$p_s'n_s = 4.50$ and $r = 5$; and, $p_s'n_s = 6.00$ and $r = 8$.

Parallel patterns run diagonally from upper left to lower right. This tendency can also be discerned in Table 2 for the larger values of $p_s'n_s$.

TABLE 2

Power of "Exact" Test of Homogeneity of
Paired Attribute Samples Proportional in Size
(one-tail test, $\alpha = 0.05$)

$p'_s n_s$	r = 1				r = 2				r = 3				r = 5				r = 8			
	p'_c/p'_s				p'_c/p'_s				p'_c/p'_s				p'_c/p'_s				p'_c/p'_s			
	1	2	3	4.5	1	2	3	4.5	1	2	3	4.5	1	2	3	4.5	1	2	3	4.5
0.75	.001	.009	.040	.139	.003	.022	.062	.144	.001	.008	.023	.065	.005	.019	.042	.082	.003	.013	.026	.059
1.125	.002	.029	.104	.286	.008	.046	.118	.269	.003	.017	.050	.133	.008	.031	.066	.137	.006	.024	.050	.100
1.50	.004	.054	.173	.433	.011	.070	.177	.372	.004	.030	.084	.217	.011	.041	.093	.180	.009	.034	.070	.136
2.25	.010	.104	.305	.653	.020	.114	.272	.529	.008	.059	.163	.375	.014	.061	.137	.285	.013	.050	.105	.210
3.00	.015	.152	.428	.790	.024	.149	.351	.650	.013	.092	.244	.513	.017	.079	.190	.372	.015	.085	.135	.280
4.50	.022	.247	.623	.935	.029	.199	.470	.804	.020	.152	.377	.696	.020	.116	.275	.539	.020	.094	.211	.413
6.00	.026	.330	.754	.981	.030	.239	.574	.901	.025	.196	.474	.809	.024	.149	.354	.653	.024	.122	.276	.521
9.00	.032	.470	.903	.998	.030	.336	.757	.978	.027	.269	.637	.933	.026	.198	.478	.811	.028	.164	.376	.675
12.00	.034	.590	.964	1.000	.030	.434	.864	.996	.028	.344	.758	.978	.026	.251	.599	.908	.028	.196	.458	.779

NOTE: $p'_c n_c$ and $p'_s n_s$ are the expected number of defective items (or defects) in samples of size n_c and $n_s = r\ n_c$, respectively.

TABLE 3

Power of "Approximate" Test of Homogeneity of Paired Attribute Samples Proportional in Size (one-tail test, $\alpha = 0.05$)

$p'_s n_s$	r = 1				r = 2				r = 3				r = 5				r = 8			
	p'_c/p'_s				p'_c/p'_s				p'_c/p'_s				p'_c/p'_s				p'_c/p'_s			
	1	2	3	4.5	1	2	3	4.5	1	2	3	4.5	1	2	3	4.5	1	2	3	4.5
0.75	.020	.097	.214	.407	.028	.098	.188	.352	.014	.050	.102	.202	.070	.136	.200	.284	.044	.089	.144	.198
1.125	.036	.157	.317	.551	.044	.145	.272	.464	.022	.080	.163	.308	.074	.149	.223	.330	.048	.101	.156	.238
1.50	.049	.199	.389	.652	.054	.182	.336	.554	.029	.109	.220	.409	.072	.151	.241	.356	.049	.109	.184	.275
2.25	.062	.252	.488	.774	.066	.232	.429	.678	.040	.159	.318	.550	.062	.154	.264	.432	.050	.123	.213	.354
3.00	.065	.285	.564	.862	.070	.264	.495	.761	.048	.198	.390	.652	.056	.165	.315	.515	.051	.144	.255	.424
4.50	.060	.342	.702	.955	.068	.307	.591	.869	.055	.249	.500	.790	.053	.191	.394	.653	.053	.172	.315	.535
6.00	.053	.408	.807	.987	.059	.341	.679	.937	.055	.294	.597	.884	.052	.231	.461	.771	.053	.197	.371	.625
9.00	.049	.538	.927	.999	.053	.422	.810	.986	.056	.384	.743	.959	.049	.277	.593	.865	.053	.226	.466	.754
12.00	.049	.648	.974	1.000	.051	.515	.903	.998	.057	.458	.834	.988	.049	.331	.678	.936	.051	.272	.553	.840

NOTE: $p'_c n_c$ and $p'_s n_s$ are the expected number of defective items (or defects) in samples of size n_c and $n_s = r\, n_c$, respectively.

Curiously, the value of the power function depends essentially upon the position of r in the Fibonacci sequence and the position of the nuisance parameter $p_s'n_s$ in one of two interpenetrating geometric series whose first terms differ by 0.375. Another way of viewing the sequence of $p_s'n_s$ values is as follows: $\frac{9}{8} \not\angle \frac{3}{8}$, $\frac{9}{4} \not\angle \frac{3}{4}$, $\frac{9}{2} \not\angle \frac{3}{2}$ and $9 \not\angle 3$.

8. COMBINATION OF TESTS OF POISSON VARIATES. When the sample size ratio r is varied from trial to trial or the classification of a defect is altered with each examination, pooling of sampling results is inappropriate for application of tests represented by (8), (17), (19), and (23). What is needed is an omnibus type test to combine all of the evidence obtained to provide a single measure of confidence in the supplier's inspection results.

From the risks associated with the "exact" and "approximate" tests under the null hypothesis we can expect a certain frequency of significant differences. Further, from the β risks associated with these tests we can expect a certain frequency of erroneous acceptances of false hypotheses. Accordingly, it is not correct to reject or accept the general hypothesis that the supplier's inspection data are as a whole unreliable as a consequence of the individual lot comparisons, which taken separately appear to yield either significant or non-significant results. The over-all test calls, therefore, for the combination of a number of independent tests of significance. Fisher ([5] , Section 21.1) has given a general method for combining the probabilities of several mutually independent tests. A number of other writers have discussed and illustrated this problem, but A. Birnbaum [21] has shown that Fisher's method is to be preferred for its somewhat more uniform sensitivity to the alternatives of interest.

The over-all test developed by Fisher deals with continuous variables. It will yield biased results if applied directly to probabilities derived from the "exact" test for Poisson variates. Lancaster [15] , David and Johnson [14] , Tocher [22] and Pearson [23] and Yates [24, 25] have considered the difficulties encountered by the combination of tests based on discontinuous variates. Since the "approximate" test and its alternates treat the number of events, d_s, d_c as continuous variates the probabilities obtained can be handled on a practical basis by application of Fisher's probability integral, which may be defined generally as follows:

Let p(x) be the probability density function of a continuous random variable x in the interval $a \leq x \leq b$, where $p(x) = 0$ for $x < a$ or $x > b$.

Then if

$$(24) \quad P = \int_a^x p(x)dx,$$

P is uniformly distributed in the interval (0,1) and $x = -2 \log_e P$ is distributed as χ^2 with 2 degrees of freedom.

If now we combine k independent probabilities, the combined probability is the product of the k separate probabilities, or

$$(25) \quad \begin{aligned} \sum (z_i) &= -2 \log_e (P_1 P_2 \ldots P_k) \\ &= -2 \sum_{i=1}^{k} \log_e P_i \end{aligned} \quad ,$$

and so has the χ^2 distribution with 2 k degrees of freedom. Thus, by means of the probability integral transformation, any number of probabilities $P_1, P_2, \ldots, P_k$ may be converted to a χ^2 value and, using the properties of the χ^2 distribution, may be summed together with the degrees of freedom to yield from published tables an over-all probability. The application of these results to continuous population is straightforward.

For discrete populations, such as the binomial represented by (5), the over-all probability is biased when the null hypothesis is true. The expectation of χ^2 for discontinuous variates is always below the theoretical value of 2. Thus, for the case $d_c + d_s = 4$ and $r = 1$, we obtain, under the null hypothesis, the binomial $(\frac{1}{2} + \frac{1}{2})^4$ and find from Table 4 below for a one-sided comparison that the expectation of $-2 \log_e P_i$ is 1.241 and the variance of the distribution is 3.527.

TABLE 4

Distribution of Probability Integral Transformation Applied to "Exact" Test for Case of Binomial $(\frac{1}{2} + \frac{1}{2})^4$ (one-sided comparison)

No. of Events		Relative Frequency	Cumulative Probability	Probability Integral Transformation
d_s	d_c	of d_s, d_c	P_i	$z_i = -2 \log_e P_i$
4	0	0.0625	1.0000	0
3	1	0.2500	0.9375	0.1291
2	2	0.3750	0.6875	0.7494
1	3	0.2500	0.3125	2.3263
0	4	0.0625	0.0625	5.5452

NOTE:	Expectation	Variance
$\chi^2_f = 2$ (theoretical)	2.000	4.000
$z_i = -2 \log_e P_i$	1.241	3.527

Similarly, for the case of the binomial $(1/3 + 2/3)^5$, which can be derived from (5) the expectation of χ^2 is 1.314 and the variance of the distribution is 2.482. There is clearly considerable bias when the probability integral transformation is applied to the probabilities derived from the "exact" test. In contrast, Table 5 below indicates comparative lack of bias in the "approximate" test (17) when we wish to combine its results for a series of independent determinations to verify a common hypothesis: that the supplier's inspection records are reliable. For the binomial distribution just discussed, where $d_t = 5$ and $p = 1/3$, Table 5 indicates that for Cox's "approximate" method the χ^2 expectation is 2.042 and the variance of the binomially-distributed probability integral transformation is 4.393. Even for an extremely small number of observed defects the continuity correction of the "approximate" test is very effective.

TABLE 5

Expectances and Variances of Binomially-Distributed Probability Integral Transformations Derived from "Approximate" Tests of Poisson Variates (one-sided comparison)

p n	$\frac{1}{1+r} = 1/2$		$\frac{1}{1+r} = 1/3$		$\frac{1}{1+r} = 1/4$		$\frac{1}{1+r} = 1/6$		$\frac{1}{1+r} = 1/9$	
d_t	$E(z_i)$	$Var(z_i)$	$E(z_i)$	$Var(z_i)$	$E(z_i)$	$Var(z_i)$	$E(z_i)$	$Var(z_i)$	$E(z_i)$	$Var(z_i)$
5	2.045	4.364	2.042	4.393	2.044	4.392	2.056	4.409	2.084	4.391
4	2.050	4.316	2.051	4.253	2.056	4.485	2.072	4.485	2.111	4.453
3	2.045	4.108	2.067	4.540	2.074	4.604	2.101	4.585	2.159	4.431
2	2.024	3.540	2.086	4.463	2.106	4.686	2.158	4.678	2.257	4.458
1	1.905	2.259	2.084	3.630	2.170	4.165	2.302	4.377	2.474	4.146

NOTES: (a) $z_i = -2 \log_e I \frac{1}{1+r} (d_c + 0.5, d_s + 0.5)$.

(b) Conditionally on d_t, z_i is binomially distributed with parameters, $\frac{1}{1+r}$ and d_t.

(c) $E (\chi^2)_{f = 2} = 2.000$

$Var(\chi^2)_{f = 2} = 4.000$

TABLE 6

Expectances and Variances of Binomially-distributed Probability Integral Transformations Derived from Median-probability Tests of Poisson Variates (one-sided comparison)

p / n	$\frac{1}{1+r} = 1/2$		$\frac{1}{1+r} = 1/3$		$\frac{1}{1+r} = 1/4$		$\frac{1}{1+r} = 1/6$		$\frac{1}{1+r} = 1/9$	
d_t	$E(z_i)$	$Var(z_i)$	$E(z_i)$	$Var(z_i)$	$E(z_i)$	$Var(z_i)$	$E(z_i)$	$Var(z_i)$	$E(z_i)$	$Var(z_i)$
5	1.940	3.461	1.930	3.444	1.927	3.417	1.910	3.332	1.889	3.240
4	1.936	3.382	1.917	3.405	1.913	3.333	1.895	3.253	1.868	3.144
3	1.940	3.225	1.913	3.302	1.897	3.242	1.871	3.125	1.841	2.982
2	1.953	2.858	1.911	3.154	1.882	3.142	1.841	3.009	1.796	2.831
1	1.981	1.975	1.940	2.553	1.898	2.754	1.829	2.814	1.754	2.693

NOTES: (a) $z_i = -\log_e \left[I_{\frac{1}{1+r}}(d_c, d_s + 1) + I_{\frac{1}{1+r}}(d_c + 1, d_s) \right]$.

(b) Conditionally on d_t, z_i is binomially distributed with parameters, $\frac{1}{1+r}$ and d_t.

(c) $E\ (\chi^2)_{f=2} = 2.000$

$Var(\chi^2)_{f=2} = 4.000$

Similarly, for the median probability test denoted by (19), the χ^2 expectations and the variances of the binomially distributed probability integral transformation (25) approach the theoretical values of 2.000 and 4.000, respectively. Table 6 discloses that for the case $d_t = 5$ and $p = 1/3$, the expectation and variance are, respectively, 1.930 and 3.444. A comparison of Table 5 and Table 6 reveals that (if d_t occurrences are few and the sample size ratio r is large) the median probability test tends to be negatively biased with variances less than the theoretical and the approximate test of Cox positively biased with variances greater than the theoretical. In product verification inspection, where the probabilities of each trial are combined, the Cox test maintains its power to detect inspection discrepancies with a small validation sample and guard against insidious differences. On the other hand, the median probability method of Lancaster provides a more conservative approach before taking serious action. When an assignable cause can be readily established for a statistically significant discrepancy, the Cox test is the method of choice.

An alternate method of combining probabilities of Poisson variates is to combine the X^2 values of tests represented by (20), or more generally by (22), according to the addition theorem for the χ^2 distribution. The sum of k values of χ^2, each with 1 degree of freedom, is distributed as χ^2 with k degrees of freedom. This test lacks power in detecting a difference that is consistently one-sided. An alternative is to compute the k values of equation (23) and add them, taking account of the signs of differences. As X is approximately normally distributed with unit standard deviation and zero mean, the sum of k X-values is approximately normally distributed with zero mean and standard deviation equal to $k^{\frac{1}{2}}$. The test function is the normal deviate

(26) $\sum X/k^{\frac{1}{2}}$.

Equations (22) and (26) were applied by the author to a set of 2 X 2 tables treated by Yates [25] . The total sample sizes of the individual trials differed by a factor of ten and the ratio of sample sizes within a trial varied from 0.55 to 1.98. The relative incidence of events were all in the range 1.8 - 9.5%. The significance levels obtained by Yates, by using conventional methods for combining probabilities, were all in close agreement with the value derived from (26).

9. MISCELLANEOUS DESIDERATA. a. Continuing Tests - Aroian and Levene [26] have pointed out that in the classical theory of testing

hypotheses a decision is made after a single trial, with the consequence that no further observations need be made. Cumulation of test results in sequential analysis or in combination of test probabilities also lead to a termination of observations. In product verification inspection, as well as in quality control work, observations are taken in sequence. At regular intervals a decision is made to follow a certain course of action. This type of test is called a continuing test since the observations are continued indefinitely.

If we suppose that in product verification inspection the supplier's sampling results are in accordance with the consumer's, there will be a probability of α at each decision point of finding a significant difference and action will be taken unnecessarily once in every $1/\alpha$ decision points in the long run. Now, if divergent inspection results should suddenly appear due to real differences in inspection practice there is a probability γ of taking action at each decision point (assuming that p_c'/p_s' is constant from trial to trial) until remedial action has been taken. Then the decision to take action will be made at say the K^{th} decision point. The probability that $K = K_o$ is the probability that we fail to take action at the first $K_o - 1$ points and do take it at the $K_o{}^{th}$, or

$$(27) \quad P\ (K = K_o) = (1 - \gamma)^{K_o - 1}\ \gamma \quad .$$

Similarly,

$$(28) \quad P\ (K \leq K_o) = 1 - (1 - \gamma)^{K_o} = \epsilon \quad .$$

If we think in terms of continuing tests, we realize that if we do not take action at the first decision point after a real discrepancy arises, we can still do so at the second, third, etc. If we fix $P\ (K \leq K_o)$ and K_o we can choose γ from (28). Then we can state that departure from inspection accordance can be detected by the K_0 decision point with a probabili·y of ϵ .

b. Poisson Approximation to the Binomial Distribution -

Many textbooks in probability have stated that if $p' \to 0$ and $n \to \infty$ so that $np' \to \lambda$ where λ is fixed and $0 < \lambda < \infty$, the binomial distribution converges to the Poisson distribution with expectation λ . If we let p_i denote the success probability of the i^{th} trial, $i = 1, 2, \ldots, n$, then the number of events which occur have the distribution sometimes called "Poisson binomial." Hodges and Le Cam [27] states that R. von Mise pointed out that the latter distribution has in the limit a Poisson distribution, provided that $n \to \infty$ and the p_i vary with n in such a way that $\sum p_i = \lambda$ is fixed and $\theta = \max \{p_1, p_2, \ldots, p_n\}$ tends to 0. The limit theorem of von Mise suggests that the Poisson approximation will be reliable provided n is large, θ is small, and λ is moderate. Hodges and Le Cam [27] showed that these requirements are unnecessarily restrictive, and that the Poisson approximation will be good provided only θ is small, whether n is small or large, and whatever value $\sum p_i$ may have. Furthermore, they state that the number of events will have approximately a Poisson distiibution even if a few of the p_i are quite large, provided these values contribute only a small part of the total $\sum p_i$.

The note of Hodges and Le Cam implies that pooling of approximate Poisson variates for an individual test of homogeneity of the sums is an appropriate procedure even though p_i values vary from trial to trial and are not all small. Further, their work supports the assumption that a test based on the Poisson distribution for comparing the number of occurrences in two samples is applicable where the probability of the attribute characteristic is small with respect to each sample. Bross and Kasten [12] have examined the practical question as to how small the proportion must be in general to yield satisfactory results with the "exact" test. They concluded that the value of 20% "would seem to serve as a rough guide for two-tailed tests at the 5% level," as compared with the usual chi-square test with Yates' correction.

Table 7 illustrates another method for investigating this question. Analogous binomial and Poisson cases, where the incidence of defectives was relatively high were used for comparison. The power of Cox's "approximate" Poisson two-sample test at the 5% level was computed for the cases selected and tabulated in Column (9) as shown. The analogous Poisson case always yields the upper bound to the power of the comparable binomial cases, but the difference in power is of no practical importance.

TABLE 7

Power of Binomial Two-sample Test Using Critical Region at 5% Level of "Approximate" Poisson Two-sample Test

(1) $p'_s n_s$	(2) $p'_c n_c$	(3) p'_s	(4) n_s	(5) p'_c	(6) n_c	(7) p'_c/p'_s	(8) n_s/n_c	(9) Power
0.75	2.25	*	*	*	*	3	1	0.214
0.75	2.25	0.075	10	0.225	10	3	1	0.203
0.75	2.25	0.05	15	0.15	15	3	1	0.206
0.75	1.125	*	*	*	*	3	2	0.188
0.75	1.125	0.075	10	0.225	5	3	2	0.175
0.75	0.75	*	*	*	*	3	3	0.102
0.75	0.75	0.05	15	0.15	5	3	3	0.090
0.75	0.45	*	*	*	*	3	5	0.200
0.75	0.45	0.03	25	0.09	5	3	5	0.200
1.50	4.50	*	*	*	*	3	1	0.389
1.50	4.50	0.15	10	0.45	10	3	1	0.373
1.50	4.50	0.06	25	0.18	25	3	1	0.386
1.50	2.25	*	*	*	*	3	2	0.336
1.50	2.25	0.15	10	0.45	5	3	2	0.326
1.50	2.25	0.05	30	0.15	15	3	2	0.335
1.50	1.50	*	*	*	*	1	1	0.049
1.50	1.50	0.15	10	0.15	10	1	1	0.038
1.50	1.50	0.06	25	0.06	25	1	1	0.045
1.50	0.75	*	*	*	*	1	2	0.054
1.50	0.75	0.15	10	0.15	5	1	2	0.042
1.50	0.75	0.05	30	0.05	15	1	2	0.051
2.25	6.75	*	*	*	*	3	1	0.489
2.25	6.75	0.15	15	0.45	15	3	1	0.489
2.25	2.25	*	*	*	*	1	1	0.062
2.25	2.25	0.15	15	0.15	15	1	1	0.050
2.25	1.35	*	*	*	*	3	5	0.264
2.25	1.35	0.09	25	0.27	5	3	5	0.238
2.25	1.35	0.045	50	0.135	10	3	5	0.259

* Not applicable. For comparison with binomial cases, computation of power given in Column (9) was based on the analogous Poisson case with distribution parameters shown in Column (1) and Column (2).

c. Small Inspection Lots - Referring to Table 1, with lot sizes assumed to be infinite, the power of the "approximate" test for binomial cases shown in Table 7 was computed from the product of two independent binomials giving $P\,(d_s,d_c \mid p_s',p_c',n_s,n_c)$ equal to

$$(29)\qquad \left\{\frac{n_s!}{d_s!(n_s-d_s)!}\;p_s'^{\;d_s}\;q_s'^{\;(n_s-d_s)}\right\}\left\{\frac{n_c!}{d_c!(n_c-d_s)!}\;p_c'^{\;d_c}\;q_c'^{\;(n_c-d_c)}\right\}.$$

Entering in (29) the critical values of the "approximate" test for Poisson variates (17) for $\alpha = 0.05$, and considering points which deviate more from the null hypothesis, absolute probabilities were summed to give the true power of the test under the assumption of infinite lot sizes. The question is whether any further loss of power occurs if samples are drawn from finite lots.

A complication ensues when samples are drawn from finite lots that are not independent. Thus in product verification sampling, where inspection lots may on occasions be small, the number of defectives removed by the first sample affects the probability of a defective in the validation sample. Moreover, the inspection practice is not to return defectives found in the sample to the lot offered for acceptance.

If two successive samples, n_s, and n_c are drawn without replacement from a finite lot of size N characterized initially by p_s' by the supplier and p_c' by the consumer, the discrepancy being merely in the count of the number of defectives in the lot, the probability getting the result d_s,d_c is

$$(30)\qquad \frac{\binom{N-p_s'N}{n_s-d_s}\binom{p_s'N}{d_s}}{\binom{N}{n_s}}\cdot\frac{\binom{N-n_s-(p_c'N-d_s)}{n_c-d_c}\binom{p_c'N-d_s}{d_c}}{\binom{N-n_s}{n_c}}$$

This expression can be used for fixed N, n_s, n_c to compute the actual probabilities for a test of homogeneity applied to a finite lot. One case was computed: $N = 160$, $n_s = 30$, $n_c = 15$, $p_s' = 0.05$ and $p_c' = 0.15$, for comparison with analogous Poisson and binomial cases. The power of the "approximate" test (17) for the jointly dependent hypergeometric distribution case is 0.409, which exceeds the comparable values in Table 7. Because of the extensive calculations required and since p_s' and p_c' are generally small so that the Poisson approximation holds, the power functions of the "approximate" test for other dependent hypergeometric distributions were not computed.

The conditional probability of getting the result d_s, d_c, under the null hypotheses of $p_o' = p_s' = p_c'$, is obtained by dividing (30) by (31), the expression for the probability of d_t:

$$(31) \qquad \frac{\binom{N-p_o'N}{n_t-d_t}\binom{p_o'N}{d_t}}{\binom{N}{n_t}}$$

which yields upon simplification

$$(32) \qquad \frac{n_s!\,n_c!\,d_t!\,(n_t-d_t)!}{d_s!\,d_c!\,(n_s-d_s)!(n_c-d_c)!\,n_t!}$$

It will be noted this representation of the conditional probability for d_s, d_c, is identical to the expression obtained by application of Fisher's "exact" test to the results of Table 1.

d. Estimating Product Quality - Results of acceptance sampling results can provide estimates of the over-all quality of the lots accepted and rejected by a sampling plan but not of the segregated portions. Consider the special case where lot quality is binomially-controlled. From a theorem by Mood [27] , if the distribution is binomial with parameter p', then the number of defectives in the samples and in the remaining part of the lots are independently and binomially distributed with the same parameter p'. We would therefore expect that the proportion defective in the remaining parts of lots rejected by an acceptance plan would be equal to the proportion defective in the remaining parts of lots accepted by an acceptance plan. But acceptance sampling results give lot quality estimates of the fraction of production rejected, in terms of proportion defective, which are generally higher than the lot quality estimates of the fraction of product accepted by the sampling plan.

This bias is evident if we consider the effect of the OC curve of a single sampling plan

$$(33) \quad Lp' = \sum_{d = o}^{c} \binom{n}{d} q'^{\,n-d} p'^{\,d} \quad ,$$

where Lp' denotes the probability of acceptance of lots produced by a binomially-controlled process with parameter p'. Assume n and c have been chosen so that $o < Lp' < 1$. All accepted lots will be characterized by $\frac{d}{n} \leq \frac{c}{n}$ and all rejected lots by $\frac{d}{n} > \frac{c}{n}$. It is clear that the estimates of lot quality afforded by the acceptance sampling results do not reflect the true quality p'.

Product verification sampling, being independent of the supplier's acceptance sampling system, can be used for evaluating the true quality of the supplies offered for consumer acceptance. Moreover, the validation sampling results serve to check the supplier's reported process average for control of reduced, normal or tightened inspection under MIL-STD 105.

10. STATISTICAL CRITERIA FOR PAIRED ATTRIBUTE SAMPLINGS. Paired attribute sampling results can be conveniently tested for statistical significance by means of tables providing critical values for the homogeneity tests described for Poisson variates. For a given number of total defectives, d_t, observed in the supplier's and consumer's samples, when α is specified, limits can be set for either d_c or d_s. This arrangement enumerates the boundary points of the critical region of the test. However, the consumer usually desires to compare his sample results, associated with the sample results recorded by the supplier, against a "rejection number."

Table I of DoD Handbook H109, included in the Appendix of this paper, was derived from (17). It sets forth an action number, depending on r, for each value of d_s, which may be recorded by the supplier. When an action number, denoted by $d_c(A)$, is reached or exceeded, the consumer adopts a course of action on the premise that a discrepancy actually exists in the supplier's inspection system. Tables IA and IB, derived from (8) and (19), respectively, are alternate sets of critical values included in the Appendix for comparison with Table I. The critical values of these tables correspond to $\alpha = 0.05$ for a one-sided test.

The probability integral transformation of Table 5 or Table 6, for a given d_s,d_c and r, can be readily evaluated from Tables of the Incomplete Beta-Function Ratio [29] and natural logarithm tables.

Table II of DoD Handbook H109, subdivided into five sections corresponding to r values 1, 2, 3, 5 and 8, respectively, yields directly the probability integral transformations, reduced by one-half, derived from "approximate" tests (17) of Poisson variates. These values, designated as "check ratings," when doubled are approximately distributed as χ^2 with 2 degrees of freedom.

Table III of DoD Handbook H109 is a modified, extended table of the percentage points of the χ^2 distribution for even-numbered degrees of freedom. As Table III is used in conjunction with Table II, the critical values tabulated are $\frac{1}{2}\chi^2_{2k}$ for 2 k degrees of freedom, where k is the number of probabilities to be combined, i.e., number of lots verified. The warning and action limits in Table III have been set at the 0.05 and 0.01 significance levels, respectively, and the median value at the 0.50 level.

The accumulation of check ratings serves to summarize all available sampling data bearing on the reliability of the supplier's inspection results. Furthermore, the ratings establish an objective degree of confidence in the relative accuracy of the supplier's results of sampling inspection. In this connection, the following graphical device may be used to show homogeneity of the paired sampling results: Using semi-logarithmic graph paper, plot the check ratings obtained from Table II of DoD Handbook H109 on the logarithmic scale against the ordinal number of the test on the arithmetic scale. Superimpose on the chart two horizontal lines corresponding to the median and 100 α% levels, respectively, of the $\chi^2/2$ value for 2 degrees of freedom. Any pronounced runs above or below the median line, or marked divergence about the 100 α% line, will indicate the likelihood of an assignable cause at work, which should be investigated.

Table IV of DoD Handbook was derived from (28) where $(1 - \gamma)$ represents the probability of acceptance of the null hypothesis of homogeneity on the basis of a single trial. The probability of acceptance of the hypothesis of homogeneity and not taking action in K trials is designated by $(1 - \epsilon)$. Table IV emphasizes the continuing nature of the test for inspection concordance and is useful for augmenting the power of a single test for homogeneity.

The power of the "approximate" test (17) for Poisson variates given in Table 3 of this paper furnished the basis for Table V of DoD Handbook H109. The latter shows the probability of the failure to reject different alternate hypotheses. This probability of a Type II error is also depicted in DoD Handbook H109 by an operating characteristics (OC) curve. The set of OC curves provided in the handbook are applicable for sample size ratios 1, 2, 3, 5 and 8 and nuisance parameters $p_s'n_s$ generally encountered in practice.

The OC curves used in conjunction with Table IV are useful in determining the size of the validation sample for a single trial. The level of significance α at which a test is to be conducted was predetermined as 0.05. The alternative that we wish to protect against and the risk that we are willing to take of making a Type II error need to be determined by the consumer. The OC curves will then show what sample size will satisfy the two conditions.

11. CONCLUDING REMARKS. The system of sampling inspection imposed upon the supplier operates to assure a product meeting specified quality standards. With the supplier's size of sample fixed by the acceptance plan, the size of the validation sample can be varied by the consumer subject to mathematical rules involving considerations of sampling risks, etc. But its adjustment can also be based on external evidence that the supplier is maintaining an acceptable quality control and inspection system, or that inspection aids are properly calibrated and used. Nevertheless, confirming data, generated by inspection of a portion of the product by the consumer, is sine qua non.

The validation sample used on an individual or skip-lot basis can furnish an estimate of the quality of product offered for acceptance. As this quality stabilizes at an acceptable level the consumer may stepwise shift to a smaller verification sample. At each stage he may consult the power function or OC curve of the significance test to determine his risks. Conversely, with sufficient statistical sophistication, the consumer can select a sample size ratio based on the power of test and the risk of not detecting an inspection discrepancy of importance within a predetermined number of trials.

These techniques can be extended to provide a cost basis for deciding upon the size of the validation sample. However, the objective of this paper and DoD Handbook H109 is to provide a system of product verification inspection which can be initiated and applied by a field inspector with the training in statistics that he already has. For this reason, the approach of classical statistics is used with predetermined levels of significance. Tables of critical values are provided to avoid computation and the techniques are simplified to the maximum extent possible, without substantial loss of power of test. Fortunately, the Poisson exponential limit is an effective approximation to the binomial and hypergeometric distributions commonly met in industrial practice.

The tests of homogeneity for Poisson variates described in this paper are sufficiently robust to have wide utility. They are appropriate in many practical applications where mass comparisons of attribute data are to be made and over-all conclusions are to be drawn. The "exact" test has been applied for this purpose to long tabulations of research attribute data (12). However, since this test does not equalize the actual size and the nominal significance level, there is a great loss of power. The "approximate" test and its alternates, by maintaining an effective level of significance, not only retain their power but may be combined for an over-all test of a common hypothesis.

The choice of the "approximate" test as the basis for product verification was determined by its applicability to a small number of events without loss of power. Empirical trials have shown this method to be practically as powerful as the randomization procedure described by Tocher [22] and Pearson [23] . In this connection, Lancaster [16] states that it is plausible to consider the median probability test function as the result of a randomization procedure carried out before the actual trial. However, since it is desirable to adopt one standard procedure where the same judgment is always made on the same data, the "approximate" procedure was selected as the method of choice for product verification inspection.

ACKNOWLEDGMENTS

The author is obliged to Mrs. Mary E. Ready for extensive aid in computing the tables appearing in this paper and in DoD Handbook H109. Thanks are also due Mrs. Marion W. Jones and Mrs. Mary E. Barstow for their assistance in performing calculations preliminary to the preparation of DoD Handbook H109.

REFERENCES

[1] Gause, G.R. (1948). "The Amount of Inspection as a Function of Control of Quality," Proceedings of the American Society for Testing Materials 48, 886-893.

[2] Quality Control and Reliability Handbook (Interim) H109, Office of the Assistant Secretary of Defense (Supply and Logistics), Washington 25, D. C., 6 May 1960, "Statistical Procedures for Determining Validity of Suppliers' Attributes Inspection," U. S. Government Printing Office, Washington 25, D. C.

[3] Military Standard 105B, 31 December 1958, "Sampling Procedures and Tables for Inspection by Attributes," U. S. Government Printing Office, Washington 25, D. C.

[4] Paulson, Edward and Wallis, W. A. (1947). "Planning and analyzing experiments for comparing two percentages," Techniques of Statistical Analysis, Statistical Research Group Columbia University, McGraw-Hill Book Co., Inc., New York, 247-265.

[5] Fisher, R. A. (1941). Statistical Methods for Research Workers, 8th Ed., Oliver & Boyd, Ltd., Edinburgh and London.

[6] Finney, D. J. (1948). "The Fisher-Yates test of significance in 2 X 2 contingency tables," Biometrika, 35, 145-156.

[7] Latscha, R. (1949). "Tests of significance in a 2 X 2 contingency table: extension of Finney's table," Biometrika, 40, 74-86.

[8] Mainland, Donald and Murray, I.M. (1952), "Tables for use in fourfold contingency tests," Science, 116, 591-594.

[9] Armsen, P. (1955). "Tables for significance tests of 2 X 2 contingency tables," Biometrika, 42, 494-505.

[10] Przyborowski, J. and Wilenski, H. (1940). "Homogeneity of results in testing samples from Poisson series," Biometrika, 31, 313-323.

[11] Barnard, G.A. (1944). "A test for homogeneity of Poisson series," Advisory Service on Statistical Method and Quality Control, Technical Report No. Q.C./R/18, Ministry of Supply, London.

[12] Bross, Irwin D.J. and Kasten, E.L. (1957). "Rapid analysis of 2 X 2 tables," Journal of the American Statistical Association, 52, 18-28.

[13] Cox, D.R. (1953). "Some simple approximate tests for Poisson Variates," Biometrika, 40, 354-360.

[14] David, F.N. and Johnson, N.L. (1950). "The probability integral when the variable is discontinuous," Biometrika, 37, 42-49.

[15] Lancaster, H.O. (1949). "The combination of probabilities arising from data in discrete distributions," Biometrika, 36, 370-382.

[16] Lancaster, H.O. (1961). "Significance Tests in Discrete Distributions," Journal of the American Statistical Association, 56, 223-234.

[17] Barton, D.E. (1958). "On the equivalence of two tests of equality of rate of occurrence in two series of events occurring randomly in time," Biometrika, 45, 267-268.

[18] Barnard, G.A. (1946). "Sequential Tests in Industrial Statistics," Journal of the Royal Statistical Society Supplement, 8, 1-21.

[19] Lancaster, H.O. (1952). "Statistical control of counting experiments," Biometrika, 39, 419-422.

[20] Birnbaum, Allan (1954). "Statistical methods for Poisson processes and exponential populations," Journal of the American Statistical Association, 49, 254-266.

[21] Birnbaum, Allan (1954). "Combining independence tests of significance," Journal of the American Statistical Association, 49, 559-574.

[22] Tocher, K.D. (1950). "Extension of the Neyman-Pearson theory of tests to discontinuous variates," Biometrika, 37, 130-144.

[23] Pearson, E.S. (1950). "On questions raised by the combination of tests based on discontinuous distributions," Biometrika, 37, 383-398.

[24] Yates, F. (1955). "The use of transformations and maximum likelihood in the analysis of quantal experiments involving two treatments," Biometrika, 42, 382-403.

[25] Yates, F. (1955). "A note on the application of the combination of probabilities test to a set of 2 X 2 tables," Biometrika, 42, 404-411.

[26] Aroian, Leo A. and Levene, Howard (1950). "The Effectiveness of Quality Control Charts," Journal of the American Statistical Association, 45, 520-529.

[27] Hodges, J. L. Jr. and Le Cam, Lucien (1960). "The Poisson approximation to the Poisson binomial distribution," Annals of Mathematical Statistics, 31, 737-740.

[28] Mood, A.M. (1943). "On the dependence of sampling inspection plans upon population distributions," Annals of Mathematical Statistics, 14, 415-425.

[29] Tables of the Incomplete Beta-Function, edited by Karl Pearson, (1934). Biometrika Office, University College, London.

TABLE I

Limits for Determining Discrepancies Between Supplier's and Consumer's Paired Attributes Sampling Inspections

(Critical Region at 5% Level of "Approximate" Poisson Two-Sample Test)

	$r = 1$	$r = 2$	$r = 3$	$r = 5$	$r = 8$
d_s	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$
0	3	2	2	1	1
1	5	3	3	2	2
2	7	4	3	3	2
3	9	5	4	3	2
4	11	6	5	3	3
5	12	7	5	4	3
6	14	8	6	4	3
7	15	9	6	5	3
8	17	9	7	5	4
9	18	10	7	5	4
10	19	11	8	6	4
11	21	12	8	6	4
12	22	12	9	6	5
13	23	13	9	7	5
14	25	14	10	7	5
15	26	14	10	7	5
16	27	15	11	7	5
17	28	16	11	8	6
18	30	16	12	8	6
19	31	17	12	8	6
20	32	18	13	9	6
21	34	18	13	9	6
22	35	19	14	9	7
23	36	20	14	9	7
24	37	20	14	10	7
25	39	21	15	10	7
26	40	22	15	10	7
27	41	22	16	11	7

TABLE I (Continued)

	$r = 1$	$r = 2$	$r = 3$	$r = 5$	$r = 8$
d_s	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$
28	42	23	16	11	8
29	43	24	17	11	8
30	45	24	17	11	8
31	46	25	18	12	8
32	47	25	18	12	8
33	48	26	18	12	9
34	49	27	19	13	9
35	51	27	19	13	9

r = Ratio of size of supplier's sample to that of the consumer's

d_s = Number of defectives (or defects) observed in the supplier's sample.

d_c = Number of defectives (or defects) observed in the consumer's sample.

$d_c(A)$ = "Action" limit for d_c. When this number is reached or exceeded in the consumer's sample, a course of action is adopted on the premise that a discrepancy does exist.

TABLE IA

Limits for Determining Discrepancies Between Supplier's and Consumer's Paired Attributes Sampling Inspections

(Critical Region at 5% Level of "Exact" Poisson Two-Sample Test)

	r = 1	r = 2	r = 3	r = 5	r = 8
d_s	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$
0	5	3	3	2	2
1	7	4	4	3	2
2	9	5	4	3	3
3	10	6	5	4	3
4	12	7	5	4	3
5	13	8	6	4	3
6	15	9	7	5	4
7	16	10	7	5	4
8	18	10	8	6	4
9	19	11	8	6	4
10	20	12	9	6	5
11	22	12	9	7	5
12	23	13	10	7	5
13	24	14	10	7	5
14	26	15	11	7	6
15	27	15	11	8	6
16	28	16	12	8	6
17	30	17	12	8	6
18	31	17	13	9	6
19	32	18	13	9	7
20	33	19	13	9	7
21	35	19	14	10	7
22	36	20	14	10	7
23	37	21	15	10	7
24	38	21	15	10	8
25	40	22	16	11	8
26	41	22	16	11	8
27	42	23	17	11	8

TABLE IA (Continued)

	$r = 1$	$r = 2$	$r = 3$	$r = 5$	$r = 8$
d_s	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$
28	43	24	17	12	8
29	45	24	17	12	8
30	46	25	18	12	9
31	47	26	18	12	9
32	48	26	19	13	9
33	49	27	19	13	9
34	51	27	20	13	9
35	52	28	20	13	10

r = Ratio of size of supplier's sample to that of the consumer's.

d_s = Number of defectives (or defects) observed in the supplier's sample.

d_c = Number of defectives (or defects) observed in the consumer's sample.

$d_c(A)$ = "Action" limit for d_c. When this number is reached or exceeded in the consumer's sample, a course of action is adopted on the premise that a discrepancy does exist.

TABLE IB

Limits for Determining Discrepancies Between Supplier's and Consumer's Paired Attributes Sampling Inspections

(Critical Region at 5% Level of Alternate "Approximate" Poisson Two-Sample Test)

	r = 1	r = 2	r = 3	r = 5	r = 8
d_s	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$
0	3*	2*	2*	2*	1*
1	6	4	3	2	2
2	8	5	4	3	2
3	9	6	4	3	3
4	11	6	5	4	3
5	12	7	5	4	3
6	14	8	6	4	3
7	15	9	7	5	4
8	17	10	7	5	4
9	18	10	8	5	4
10	19	11	8	6	4
11	21	12	9	6	4
12	22	12	9	6	5
13	23	13	10	7	5
14	25	14	10	7	5
15	26	14	11	7	5
16	27	15	11	8	5
17	29	16	11	8	6
18	30	17	12	8	6
19	31	17	12	8	6
20	32	18	13	9	6
21	34	18	13	9	6
22	35	19	14	9	7
23	36	20	14	10	7
24	37	20	15	10	7
25	39	21	15	10	7
26	40	22	15	10	7
27	41	22	16	11	8

TABLE IB (Continued)

	r = 1	r = 2	r = 3	r = 5	r = 8
d_s	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$	$d_c(A)$
28	42	23	16	11	8
29	44	24	17	11	8
30	45	24	17	12	8
31	46	25	18	12	8
32	47	26	18	12	8
33	48	26	19	12	9
34	50	27	19	13	9
35	51	27	19	13	9

* Critical value derived from mid-probability test function defined in Lancaster [15] .

r = Ratio of size of supplier's sample to that of the consumer's.

d_s = Number of defectives (or defects) observed in the supplier's sample.

d_c = Number of defectives (or defects) observed in the consumer's sample.

$d_c(A)$ = "Action" limit for d_c. When this number is reached or exceeded in the consumer's sample, a course of action is adopted on the premise that a discrepancy does exist.